Regresní přímka

Model:

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times k}\boldsymbol{\beta}_{k\times 1} + \mathbf{e}_{n\times 1}$$

Rovnice pro odhad β:

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

$$\widehat{\boldsymbol{\beta}} = \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Pro přímku ($k = 2$) máme

$$Y_i = \beta_0 + \beta_1 x_i \quad i = 1, \ldots, n$$

Pro přímku můžeme vektor $\mathbf{Y}$, matici plánů $\mathbf{X}$ a vektor $\boldsymbol{\beta}$ zapsat následovně:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

Tedy odtud máme:

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

$$\det \mathbf{X}'\mathbf{X} = n\sum x_i^2 - (\sum x_i)^2 = n\sum (x_i - \overline{x})^2$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \frac{\sum x_i^2}{n\sum x_i^2 - (\sum x_i)^2} & \frac{-\sum x_i}{n\sum x_i^2 - (\sum x_i)^2} \\ \frac{-\sum x_i}{n\sum x_i^2 - (\sum x_i)^2} & \frac{n}{n\sum x_i^2 - (\sum x_i)^2} \end{pmatrix}$$

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \end{pmatrix}$$

$$\mathbf{b} = \begin{pmatrix} \frac{\sum x_i^2}{n\sum x_i^2 - (\sum x_i)^2} \sum Y_i + \frac{-\sum x_i}{n\sum x_i^2 - (\sum x_i)^2} \sum x_i Y_i \\ \frac{-\sum x_i}{n\sum x_i^2 - (\sum x_i)^2} \sum Y_i + \frac{n}{n\sum x_i^2 - (\sum x_i)^2} \sum x_i Y_i \end{pmatrix}$$

Odhady regresních koeficientů pro přímku jsou tedy:

$$b_1 = \frac{n\sum x_i Y_i - \sum x_i \sum Y_i}{n\sum x_i^2 - (\sum x_i)^2}$$
$$b_0 = \frac{\sum x_i^2 \sum Y_i - \sum x_i \sum x_i Y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{1}{n}(\sum Y_i - b_1 \sum x_i)$$

S využitím vztahů

$$\begin{aligned} s_x^2 &= \frac{1}{n-1}\sum (x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum x_i^2 - n\overline{x}^2\right) \\ &= \frac{1}{n-1}\left[\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right] \\ s_{xy} &= \frac{1}{n-1}\sum (x_i - \overline{x})(Y_i - \overline{Y}) = \frac{1}{n-1}\left(\sum x_i Y_i - n\overline{x}\overline{Y}\right) \\ &= \frac{1}{n-1}\left(\sum x_i Y_i - \frac{1}{n}\sum x_i \sum Y_i\right) \end{aligned}$$

Můžeme tyto vzorce přepsat pomocí výběrových charakteristik:

$$b_1 = \frac{s_{xy}}{s_x^2} \quad b_0 = \overline{Y} - \frac{s_{xy}}{s_x^2}\overline{x}$$

Dosazením b_0 a b_1 za β_0 a β_1 dostaneme:

$$\widehat{Y} = b_0 + b_1 x = \overline{Y} + \frac{s_{xy}}{s_x^2}(x - \overline{x}) = \overline{Y} + r_{xy}\frac{s_y}{s_x}(x - \overline{x})$$

Pro reziduální rozptyl platí maticový zápis

$$S_e = (\mathbf{Y} - \widehat{\mathbf{Y}})'(\mathbf{Y} - \widehat{\mathbf{Y}}) = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$$

rozepsáním dostaneme

$$S_e = \sum (Y_i - \widehat{Y}_i)^2 = \sum (Y_i - b_0 - b_1 x_i)^2 = \sum Y_i^2 - b_0 \sum Y_i - b_1 \sum x_i Y_i$$

Normalita

Pokud platí

$$\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

pak

$$\mathbf{b} \sim N_k(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

Nestranným odhadem pro parametr σ^2 je veličina $s^2 = \frac{S_e}{n-k}$.

Dále platí

$$T_i = \frac{b_i - \beta_i}{\sqrt{s^2 v_{ii}}} \sim t(n-k)$$

kde v_{ii} je $(\mathbf{X}'\mathbf{X})^{-1}_{ii}$

Pro přímku ($k = 2$) máme

$$v_{00} = \frac{\sum x_i^2}{n\sum x_i^2 - (\sum x_i)^2} = \frac{\sum x_i^2}{n(n-1)s_x^2}$$
$$v_{11} = \frac{n}{n\sum x_i^2 - (\sum x_i)^2} = \frac{1}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}$$

Dosazením za v_{ii} ($i = 0,\ 1$) dostaneme

a) pro β_0

$$T_0 = \frac{b_0 - \beta_0}{s\sqrt{\sum x_i^2}}\sqrt{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} = \frac{b_1 - \beta_1}{s}\left(s_x\sqrt{\frac{n(n-1)}{\sum x_i^2}}\right) \sim t(n-2)$$

Odtud dostáváme interval spolehlivosti pro parametr β_0

$$\left(b_0 - t_{1-\frac{\alpha}{2}}(n-2)\frac{s}{s_x}\sqrt{\frac{\sum x_i^2}{n(n-1)}}\ ;\ b_0 + t_{1-\frac{\alpha}{2}}(n-2)\frac{s}{s_x}\sqrt{\frac{\sum x_i^2}{n(n-1)}}\right)$$

b) pro β_1

$$T_1 = \frac{b_1 - \beta_1}{s}\sqrt{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} = \frac{b_1 - \beta_1}{s}\left(s_x\sqrt{n-1}\right) \sim t(n-2)$$

Odtud dostáváme interval spolehlivosti pro parametr β_1

$$\left(b_1 - t_{1-\frac{\alpha}{2}}(n-2)\frac{s}{s_x\sqrt{n-1}}\ ;\ b_1 + t_{1-\frac{\alpha}{2}}(n-2)\frac{s}{s_x\sqrt{n-1}}\right)$$

Pro parametrickou funkci $\gamma = \mathbf{c}'\boldsymbol{\beta}$ platí

$$T = \frac{\mathbf{c}'\mathbf{b} - \gamma}{s\sqrt{\mathbf{c}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{c}}} \sim t(n-k)$$

Často máme dánu hodnotu vysvětlující proměnné x a máme odhadnout střední hodnotu vysvětlované proměnné $\mathrm{E}Y = \beta_0 + \beta_1 x = \gamma$ (pro přímku tj. $k = 2$) a testovat hypotézu $H_0 : \mathrm{E}Y = \gamma_0$ proti $H_1 : \mathrm{E}Y \neq \gamma_0$. Vyjdeme z předchozího vzorce, kde položíme $\mathbf{c}' = \begin{pmatrix} 1 & x \end{pmatrix}$. Platí

$$\mathbf{c}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{c} = \frac{1}{n} + \frac{(x-\overline{x})^2}{\sum x_i^2 - n\overline{x}^2} = \frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}$$

Tedy k testování H_0 použijeme statistiku

$$T_0 = \frac{b_0 + b_1 x - \gamma_0}{s\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}}} \sim t(n-2)$$

a příslušný interval spolehlivosti

$$\left(b_0 + b_1 x - t_{1-\frac{\alpha}{2}}(n-2)s\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}} \; ; \; b_0 + b_1 x + t_{1-\frac{\alpha}{2}}(n-2)s\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}} \right)$$

Pro $x = \overline{x}$ je šířka intervalu nejmenší. Jak se x vzdaluje od $\overline{x}$ roste šířka intervalu.

Tento interval je třeba interpretovat bodově! Pás spolehlivosti pro celou regresní přímku získáme nahrazením kvantilu $t_{1-\frac{\alpha}{2}}(n-2)$ v předchozím vzorci číslem $[2F_{1-\frac{\alpha}{2}}(2, n-2)]^{\frac{1}{2}}$.

Pro inverzi matice $\mathbf{A}_{2x2}$ platí:

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \det \mathbf{A} = ad - bc \quad \mathbf{A}^{-1} = \begin{pmatrix} \frac{d}{ad-bc} & \frac{-b}{ad-bc} \\ \frac{-c}{ad-bc} & \frac{a}{ad-bc} \end{pmatrix}$$